

Anotační manuál – pokyny pro anotační testy

Úkoly






Úkoly jsou vytvořeny tak, že představují typickou práci, kterou budete jako anotátoři muset zvládat. Nejdůležitější úkol je úkol číslo 1. Ostatní jsou doplňkové úkoly, které Vás také v omezené míře také potkají.

1. Anotace 40 vět z oblasti informací o počasí
2. Tvorba anotačního schématu dle dané databáze
3. Vymyšlení několika tématických vět
4. Lematizace

Úkol č. 1

Proveďte anotaci zadaných vět. Věty se nacházejí v souboru CD-ROM: /data/vstup.xml

Postup:

1. Zkopírujte adresář editor-annotaci z CD na pevný disk (u nás v laboratoři do E:/Users/Anotace/Vaše_příjmení)
2. Zkopírujte soubor se vstupními větami z CD-ROM do adresáře na pevný disk (u nás v laboratoři do E:/Users/Anotace/Vaše_příjmení/data).
3. V adresáři editor-annotaci na pevném disku spusťte soubor spustit.bat.
4. Pokud se program nespustí, nainstalujte Javu (nachází se na CD v adresáři Java). Pokud se program spustí, pokračujte dál.
5. V programu otevřete vstupní soubor. Protože budete dělat novou anotaci (a ne upravovat/doplňovat již rozdělanou práci), tak zvolíte příkaz nový z hlavního nástrojového panelu. V dialogovém okně poté naleznete a zvolíte vstupní soubor vstup.xml.
6. Proveďte anotaci prvních 40-ti vět. Anotaci provádíte tak, že v hlavním okně (velká bílá plocha) kliknete pravým tlačítkem a zvolíte téma (). Na téma pak znova kliknete pravým tlačítkem a zvolíte pod-téma (). Takto pokračujete tak dlouho dokud se nedostanete k lexikální třídě (tu poznáte podle této ikony ). V lexikální třídě ještě označíte slovo nebo skupinu slov, která tuto třídu tvoří. Jako slova lexikální třídy označujeme pouze významová slova (Praha, 6:00, šest hodin), nikoliv slova pomocná (tedy v, na, do, bych, ...).
7. Až dokončíte anotaci, přejdeme na další větu (). Tímto postupem anotujete prvních 60 vět vstupního souboru.
8. Nakonec je třeba anotace uložit. To provedete stisknutím tlačítka uložit (). Soubor pojmenujte pocasi-test.xml.
9. Ukončete program.

Pro anotace budete potřebovat anotační schémata (příloha 1) a jejich popis (příloha 2).

Úkol č. 2

Vytvořte anotační schémata pro zadanou databázi. Databáze se nachází na webových stránkách: <http://www.ubytovani.cz/>.

Postup:

1. Otevřete internetové stránky <http://www.ubytovani.cz/>
2. Prohlédněte si parametry vyhledávání.
3. Rukou na papír nakreslete anotační schéma. Můžete se nechat inspirovat přílohou 1 a 2.

Úkol č. 3

K anotačnímu schématu z úkolu 2 vytvořte 20 typických dotazů na ubytování (např. Kde se můžu ubytovat v chatě na Šumavě. Hledám levný hotel v Plzni.). Věty vytvořte v textovém editoru (třeba poznámkový blok, nikoliv však word). Věty uložte do souboru `ubytovani.txt`.

Úkol č. 4

Převeďte následující slova / slovní spojení do základního tvaru (tzv. lemma). Např. Plzní → Plzeň, Hradci Králové → Hradec Králové, uběhnuvše → uběhnout, bejt → být,

Postup:

1. Zkopírujte na disk soubor CD-ROM: `/data/lemmatizace.txt`
2. Otevřete tento soubor v textovém editoru (např. poznámkový blok)
3. Do souboru doplňte lemma jednotlivých slov

Soubor má následující strukturu

1. řádka: Text věty, kde se slovo pro lemmatizaci vyskytuje
2. řádka: Slovo určené pro lemmatizaci
3. řádka: Lemma zadaného slova
4. prázdná řádka

Takto se záznamy v souboru opakují

Např. (před doplněním)

(po doplnění)

Kolik je nového sněhu na Klínovci ?
Klínovci

Kolik je nového sněhu na Klínovci ?
Klínovci

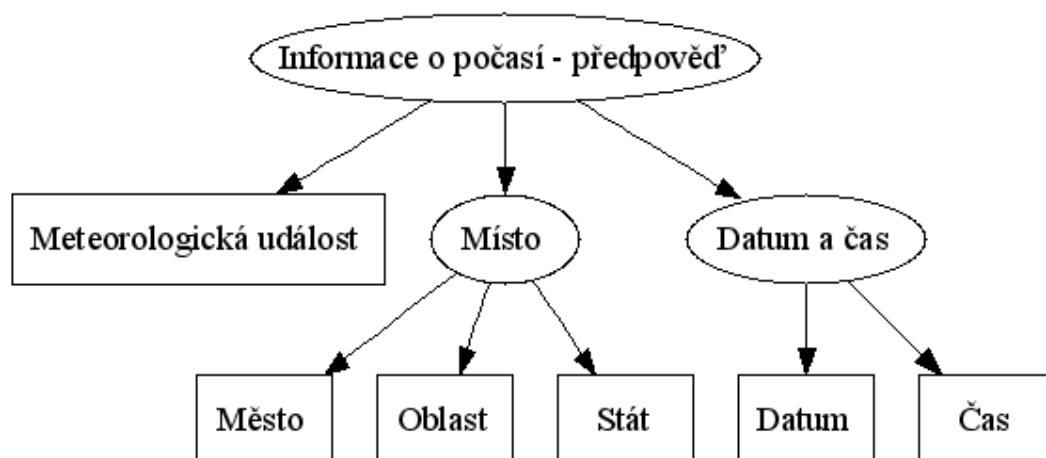
Klínovec

Další věta ...

Další věta ...

Příloha 1 – Anotační schémata

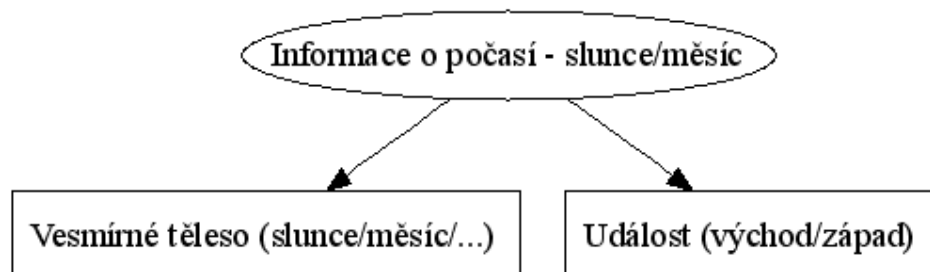
A:



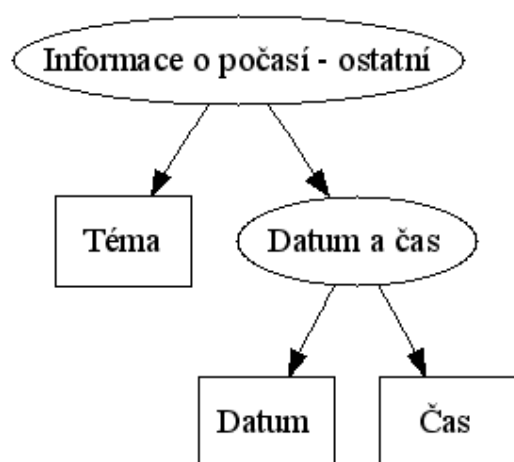
B:



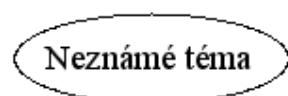
C:



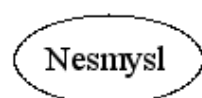
D:



E:



F:



Příloha 2 – Anotační schémata

Schéma A: Dotaz na předpověď o počasí. Uživatel se typicky ptá, jak bude v daný čas na daném místě. Druhý typ dotazu bývá konkrétnější dotaz na nějaký jev. Tedy zda bude hezky, pršet atp.

Typické věty:

- Jaké bude počasí o víkendu ?
- Jaká je předpověď počasí na příští týden ?
- Bude zítra sněžit ?

Schéma B: Dotaz na stav sněhu na lyžařských areálech. Uživatel se ptá, kolik je sněhu v nějaké oblasti.

Typické věty:

- Kolik je sněhu na Špičáku ?
- Dá se lyžovat v Jeseníkách ?

Schéma C: Žádost o informaci ohledně západů a východů slunce a měsíce. Popř. zda je vidět v noci nějaká planeta atp.

Typické věty:

- V kolik hodin zapadá slunce ?
- Kdy bude příští úplňek ?

Schéma D: Další dotazy týkající se počasí, které nespádají do předchozích témat.

Typické věty:

- Kolik stupňů má Lipno ?
- Kde je v Plzni hydrometeorologická stanice ?
- Je Vltava sjízdná ?

Schéma E: Věta dává smysl, ale nezapadá do žádného z anotačních schémat. Věta je počítačově zodpověditelná.

Typické věty:

- Kdy je otevřeno Pivovarské muzeum ?
- Je v blízkosti Plzně kemp?

Schéma F: Věta nedává smysl nebo je počítačově nezodpověditelná. Větu nemá smysl zpracovávat.

Typické věty:

- Čemu se říká babí léto?
- Jak vzniká hurikán?
- Ty kapky se dělají v lamináru ?